De vrijheidsgradencorrectie van R2

den Butter, F.A.G.; Van de Gevel, F.J.J.S.

published in
VVS Bulletin
1979

DE VRIJHEIDSGRADENCORRECTIE VAN R^2

Met toestemming van de redactie overgenomen uit de rubriek "Uit de statistische consultatie" in: *VVS Bulletin*, 12, nr. 1 (januari 1979), pp. 23-30, 22.

de vrijheidsgradencorrectie van R^2

door F.A.G. den Butter en F.J.J.S. van de Gevel*

Inleiding

Het is gebruikelijk bij regressieresultaten de correlatiecoëfficiënt R of de determinatiecoëfficiënt R^2 te vermelden. Deze grootheden geven een indruk van de kwaliteit van de aanpassing die men met de regressie heeft bereikt. Vaak is deze R of R^2 zgn. voor vrijheidsgraden gecorrigeerd. Meestal gebruikt men hiervoor dan resp. de symbolen $\bar{R}$ en $\bar{R}^2$. De filosofie achter deze vrijheidsgradencorrectie is, globaal gesproken, dat men bij de beoordeling van de aanpassingskwaliteit van de regressievergelijking dient te beseffen dat een groot aantal verklarende variabelen een bepaalde grootheid veelal beter weet te verklaren dan een klein aantal. Iedere extra verklarende variabele betekent het verlies van een vrijheidsgraad [1]. De vrijheidsgradencorrectie houdt hiermee rekening.

In dit artikel gaan we wat dieper in op het gebruik om de R^2 voor vrijheidsgraden te corrigeren. Eerst memoreren we in het kort welke betekenis de R^2 in statistische zin in de regressie-analyse heeft. Daarna wordt de wijze van corrigeren besproken en komen enkele eigenschappen van de $\bar{R}^2$ aan de orde en de verschillen die er in dat opzicht met de ongecorrigeerde R^2 zijn. In de praktijk wordt met name naar de R^2 gekeken om verschillende regressie-uitkomsten qua aanpassingskwaliteit met elkaar te kunnen vergelijken. In feite is dit alleen zinvol indien de te verklaren variabele en de waarnemingen in iedere vergelijking dezelfde zijn. Van belang is dan hoeveel een bepaalde verklarende variabele tot de R^2 bijdraagt. Deze bijdrage verschilt natuurlijk al naar gelang men de ongecorrigeerde R^2 of de gecorrigeerde $\bar{R}^2$ beschouwt. Theil (1971) [2] heeft een schema opgesteld waarin de afzonderlijke bijdragen van de verklarende variabelen in de totale verklaring worden vermeld. We bespreken enkele consequenties indien in dit schema de afzonderlijke bijdragen direct voor vrijheidsgraden gecorrigeerd worden. Tot slot van dit artikel vatten we onze bevindingen samen.

*) De auteurs zijn medewerkers van de sectie Econometrie en Wetenschappelijk Onderzoek van De Nederlandsche Bank N.V.

1) Voor een heldere en eenvoudige verhandeling over het begrip vrijheidsgraden zie J.H.C. Lisman (1977).

2) Eén van de nuttige suggesties van de redakteuren betreft een verwijzing naar een alternatief schema van E. Pedhazur (1975), dat wij vanwege onze oriëntatie op economisch onderzoek over het hoofd hebben gezien.

De vrijheidsgradencorrectie van de R^2 is uit theoretisch oogpunt een nogal ondankbaar onderwerp. Vandaar dat er in de statistische literatuur vrij weinig over geschreven is. Toch rechtvaardigt het feit dat deze correctie in de praktijk veelvuldig doch stilzwijgend wordt toegepast onze aandacht voor dit onderwerp. We hebben gepoogd op informele wijze enkele aspecten van deze vrijheidsgradencorrectie te belichten zonder daarbij al te diep in details te treden. Ons verhaal is vooral bedoeld voor de producenten en consumenten van empirisch (economisch) onderzoek en dient niet om verslag te doen van eigen, schokkende, innovaties op dit onderzoekgebied.

De R^2

De R^2 geeft aan in hoeverre men in een regressievergelijking geslaagd is de variatie van de afhankelijke variabele te verklaren. We beschouwen de regressievergelijking

$$Y_i = b_1 + b_2 X_{2i} + \ldots + b_k X_{ki} + e_i \qquad (1)$$

met Y_i de waarde van de te verklaren variabele bij waarneming i, b_1 t/m b_k de geschatte waarden van de coëfficiënten, X_{2i} t/m X_{ki} de waarden van de verklarende variabelen bij waarneming i en e_i de waarde van het residu bij die waarneming. In een steekproef met n waarnemingen is de variatie van de te verklaren variabele gelijk aan $\sum_{i=1}^{n} y_i^2$ waarbij y_i de waarde van de te verklaren variabele in afwijking van zijn gemiddelde is: $y_i = Y_i - \sum_{i=1}^{n} Y_i/n$. Van deze variatie heeft men $\sum_{i=1}^{n} e_i^2$, de kwadratensom van de residuen, niet weten te verklaren en dus $\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} e_i^2$ wel. De R^2 is gelijk aan het aandeel in de totale variatie dat door de regressievergelijking wordt verklaard (de sommatie-indices laten we verder achterwege):

$$R^2 = 1 - \frac{\Sigma e^2}{\Sigma y^2}.$$

We zijn ervan uitgegaan dat de regressievergelijking een constante term bevat. Indien dit niet het geval is, wordt de te verklaren variabele niet in afwijking van zijn gemiddelde genomen en komt in de bovenstaande formule ΣY^2 i.p.v. Σy^2 te staan. In het vervolg beschouwen we echter alleen regressievergelijkingen met een constante term.

De formule voor R^2 laat zien hoe deze grootheid berekend wordt uit de steekproefwaarnemingen. Men kan daarbij de R^2 opvatten als de schatting van een onbekende maar gedefinieerde parameter P^2, net zoals b_1 t/m b_k schattingen zijn van onbekende parameters β_1 t/m β_k uit het regressiemodel. Op deze manier is het gebruik van de R^2 te onderbouwen met een statistische theorie en zijn er waarschijnlijkheidsuitspraken over deze grootheid te doen.

De vraag is echter welke veronderstellingen men over de stochastiek in het model maakt. Zo kunnen we het regressiemodel als volgt opschrijven

$$\underline{Y}_i = \beta_1 + \beta_2 \underline{X}_{2i} + \ldots + \beta_k \underline{X}_{ki} + \underline{u}_i \qquad (2)$$

Hierbij geven we kansvariabelen of stochasten door onderstreping aan. In de bovenstaande vergelijking worden zowel de storingsterm $\underline{u}_i$ (onderling onafhankelijk, constante variantie, verwachting nul) als ook de verklarende variabelen $\underline{X}_{2i}$ t/m $\underline{X}_{ki}$ als stochasten, d.w.z. als trekkingen uit bepaalde kansverdelingen beschouwd. De Y_i en de X_{2i} t/m X_{ki} in vergelijking (1) geven de uitkomsten van dit kansproces bij waarneming i weer. In het model (2) is de P^2 een parameter uit een multivariate kansverdeling, die de correlatie tussen $\underline{Y}$ en de $\underline{X}_2$ t/m $\underline{X}_k$ beschrijft.

In economische toepassingen van de regressierekening worden de verklarende variabelen meestal niet als stochasten beschouwd, maar wordt de statistische analyse uitgevoerd gegeven de waarden van de verklarende variabelen. Het regressiemodel wordt in dat geval

$$\underline{Y}_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \underline{u}_i \qquad (3)$$

Nu is alleen de storing een trekking uit een kansverdeling. In dit geval is de definitie van de parameter P^2 minder eenvoudig en dient het verloop van de verklarende variabelen aan bepaalde voorwaarden te voldoen.

Wanneer men waarschijnlijkheidsuitspraken over de gevonden waarden van $\underline{R}^2$ wenst te doen is het van belang de kansverdeling van de schatter [3] $\underline{R}^2$ van P^2 te kennen. Indien in model (2) zowel de verklarende variabelen als ook de storingen normaal verdeeld zijn is P^2 een parameter uit een multivariate normale verdeling. Voor dit geval geeft Anderson (1958, blz. 89–96) enige alternatieven voor de verdeling van $\underline{R}^2$. In model (3) met niet-stochastische verklarende variabelen en normaal verdeelde storingen is de $\underline{R}^2$ volgens een zgn. niet centrale β-verdeling verdeeld. De Haan en Taconis-Haantjes (1978) hebben aangetoond dat hierbij $\sqrt{n}\,(\underline{R}^2 - P^2)$ in de limiet ($n \to \infty$) benaderd kan worden door een normale verdeling. Indien men echter niet het limiet-geval beschouwt, hangt de verdeling van $\underline{R}^2$ in belangrijke mate af van het verloop van de verklarende variabelen, van de variantie van de storingen en van de waarden β_2 t/m β_k in het model. Koerts en Abrahamse (1969, hoofdstuk 5) hebben voor enkele numerieke voorbeelden de invloed van deze grootheden op de verdeling van $\underline{R}^2$ onderzocht.

In de praktijk wordt vanwege al deze complicaties de verdeling van $\underline{R}^2$ bij de verslaggeving van empirisch onderzoek echter nooit gebruikt. Wel wordt veelal in standaard-regressieprogramma's de waarde van de zgn. F-toets afgedrukt. Uit de variantie-analyse geldt nl. dat

$$\frac{\underline{R}^2 / (k - 1)}{(1 - \underline{R}^2) / (n - k)} \qquad (4)$$

F-verdeeld is met $(k - 1)$ en $(n - k)$ vrijheidsgraden. Dit is een eenvoudige manier om de hypothese te toetsen dat er geen samenhang tussen Y en de verklarende variabelen bestaat (β_2 t/m $\beta_k = 0$). In feite komt dit overeen met de toets dat $P^2 = 0$.

De $\bar{R}^2$

Het gebruik van de R^2 in de regressierekening blijft echter niet beperkt tot het toetsen van de hypothese $P^2 = 0$. In dat geval zou men beter alleen de F-waarde bij de regressieresultaten kunnen vermelden en de R^2 achterwege laten. Men vertoont de R^2 om een indruk te geven van de mate van verklaring die men met de regressievergelijking heeft verkregen. Impliciet wordt daarbij vaak een vergelijking gemaakt met alternatieve regressies. Eigenlijk zou men zo'n vergelijking expliciet moeten maken door de hypothese $P_1^2 = P_2^2$ (met P_1 de P bij de eerste en P_2 de P bij de tweede regressie) te toetsen maar dit levert grote praktische moeilijkheden op. Vandaar dat men zonder een diepere statistische bijbedoeling naar de R^2 als maat van verklaring kijkt.

Formule (4) laat zien hoe bij een statistische toets voor de P^2 de vrijheidsgraden een rol spelen. Ook bij het niet statistische gebruik van de R^2 zijn deze vrijheidsgraden van belang en bestaat er aanleiding om de R^2 voor vrijheidsgraden te corrigeren. Immers

[3] Een schatter is een statistische grootheid en een functie van stochasten: in dit geval $\underline{R}^2 = 1 - \Sigma \underline{e}^2 / \Sigma \underline{y}^2$. Een schatting is een bij zo'n schatter horende realisatie.

wanneer men twee regressievergelijkingen (met dezelfde te verklaren variabele) met elkaar vergelijkt zal de vergelijking met het grootste aantal verklarende variabelen over het algemeen een hogere R^2 te zien geven. Dit is zelfs altijd zo indien die vergelijking alle verklarenden van de andere vergelijking bevat plus nog één of meer extra variabelen. Het is intuïtief begrijpelijk dat men hiervoor dient te corrigeren.

De feitelijke vrijheidsgradencorrectie van R^2 berust op de volgende gedachte. Men zou in de formule

$$R^2 = 1 - \frac{\Sigma e^2 / n}{\Sigma y^2 / n}$$

$\Sigma e^2 / n$ als een schatting voor de variantie van de storingen kunnen beschouwen en $\Sigma y^2 / n$ als een schatting voor de variantie van de te verklaren variabele. De desbetreffende schatters zijn echter niet zuiver d.w.z. de mathematische verwachting van de schatters is niet gelijk aan de waarde van de te schatten parameter. In het geval van zuiverheid dient Σe^2 door $n-k$ gedeeld te worden en Σy^2 door $n-1$, waarbij men ervan uitgaat dat de y waarnemingen zijn uit een verdeling met gelijke verwachting voor iedere trekking en dat y dus niet gegenereerd wordt door een regressiemodel met niet-stochastische verklarende variabelen. De op deze wijze voor vrijheidsgraden gecorrigeerde $\bar{R}^2$ wordt

$$\bar{R}^2 = 1 - \frac{\Sigma e^2 / (n-k)}{\Sigma y^2 / (n-1)} = R^2 - \frac{k-1}{n-k}(1-R^2) .$$

Deze formule laat zien dat de gecorrigeerde $\bar{R}^2$ altijd kleiner is dan de ongecorrigeerde R^2 en dat de correctiefactor groter wordt naarmate het aantal verklarende variabelen groter is, naarmate het aantal waarnemingen kleiner is en naarmate de R^2 lager is.

Ofschoon zuiverheid een motief is om de R^2 voor vrijheidsgraden te corrigeren verschaft de $\bar{R}^2$ ons geen zuivere schatting voor P^2. In het multivariate normale model (2) zou men, om een (vrijwel) zuivere schatting te krijgen, de R^2 als volgt moeten corrigeren[4)]

$$\tilde{R}^2 = R^2 - \frac{(1-R^2)}{n}(k-1-2R^2)$$

en in het voor economische toepassingen meer relevante model (3) heeft Barten (1962)[5)] laten zien dat bij normaal verdeelde storingen de correctie

$$\hat{R}^2 = R^2 - \frac{(1-R^2)}{n}[k-(1-R^2)(1+2R^2)]$$

bij benadering een zuivere schatting oplevert. Het verschil tussen de voor vrijheidsgraden gecorrigeerde $\bar{R}^2$ en deze laatste correctie is echter nogal klein. Bovendien worden de beide alternatieve correcties in de praktijk vrijwel nooit toegepast[6)]. Daarom laten we

4) Zie hiervoor H. Theil (1971, blz. 190- 191).

5) De Haan en Taconis-Haantjes (1978) geven de (scherpe) voorwaarden met betrekking tot de verklarende variabelen aan, waarbij deze formule van Barten geldt.

6) Een uitzondering hierop vormt Driehuis, die bij de regressieresultaten van zijn kwartaalmodel de met de formule van Barten gecorrigeerde determinatiecoëfficiënt vermeldt. Zie W. Driehuis (1972, blz. 24).

deze hier verder rusten en richten onze aandacht uitsluitend op de voor vrijheidsgraden gecorrigeerde determinatiecoëfficiënt.

Verschillen tussen R^2 en $\bar{R}^2$

De waarde van de parameter P^2 ligt, zowel in model (2) als in model (3) altijd tussen 0 en 1. Hetzelfde geldt voor de ongecorrigeerde R^2. De gecorrigeerde $\bar{R}^2$ kan daarentegen in bepaalde gevallen een negatieve waarde aannemen en dit is natuurlijk een onaangename eigenschap, aangezien het om de schatting van een parameter gaat die altijd positief is. Anderzijds is dit niet zo'n ernstig bezwaar van de gecorrigeerde $\bar{R}^2$, want indien deze een negatieve waarde krijgt, is er toch al nauwelijks sprake van enige correlatie. In dat geval is de ongecorrigeerde R^2 altijd kleiner dan $(k-1)/(n-1)$. Bij zeg 20 waarnemingen en 3 verklarende variabelen betekent het dat de R^2 kleiner dan 0,158 is wanneer men op een negatieve $\bar{R}^2$ uitkomt.

Het kernpunt van de vrijheidsgradencorrectie ligt bij het vergelijkbaar maken van de R^2 bij specificaties met een verschillend aantal verklarende variabelen, maar dezelfde te verklaren variabele en dezelfde waarnemingen. Indien nl. de te verklaren variabele en/of de waarnemingen ook verschillen, is een vergelijking van de regressievergelijkingen met R^2 (of $\bar{R}^2$) als maatstaf niet erg opportuun. Daarom zullen we nagaan wat er gebeurt met de R^2 en de $\bar{R}^2$ indien men bij een bepaalde regressie een extra verklarende variabele toevoegt. Zij R_h^2 de determinatiecoëfficiënt van de regressie die de h^e verklarende variabele niet bevat en R^2 de determinatiecoëfficiënt van de regressie waaraan deze h^e variabele is toegevoegd. Omdat de kleinste kwadratenmethode de kwadratensom van de residuen minimaliseert, en dus R^2 maximaliseert, geldt $R^2 \geqslant R_h^2$ d.w.z. de ongecorrigeerde R^2 wordt bij toevoeging van een extra verklarende variabele nooit kleiner. Dit geldt echter niet voor de gecorrigeerde $\bar{R}^2$. Indien $\bar{R}^2$ kleiner is dan $\bar{R}_h^2$ en de gecorrigeerde $\bar{R}^2$ dus afneemt bij toevoeging van een verklarende variabele, weegt de extra verklaring die deze variabele toevoegt niet op tegen het verlies van één vrijheidsgraad dat men door het opnemen van deze variabele in de regressie lijdt. Uit de formule van de vrijheidsgradencorrectie is gemakkelijk te berekenen dat de gecorrigeerde $\bar{R}^2$ gelijk blijft wanneer de ongecorrigeerde R^2 met $(1-R^2)/(n-k)$ toeneemt d.w.z.

$$\text{indien } R^2 - R_h^2 = \frac{1-R^2}{n-k} \quad \text{dan} \quad \bar{R}^2 - \bar{R}_h^2 = 0 .$$

Vaak worden bij regressieresultaten de zgn. t-waarden van de coëfficiënten vermeld. Met zo'n t-waarde kan de hypothese getoetst worden of de desbetreffende coëfficiënt van nul verschilt. Het is gebruikelijk om de waarde 2 voor deze grootheid bij een voldoend aantal vrijheidsgraden als kritische grens aan te houden. Bij een (absolute) waarde van de t-toets groter dan 2 is de coëfficiënt "significant" en bij een t-waarde kleiner dan 2 concludeert de onderzoeker, veelal enigermate teleurgesteld, dat de verklarende variabele waarop de coëfficiënt betrekking heeft de te verklaren variabele niet duidelijk beïnvloedt.

Er bestaat het volgende verband tussen de t-waarde van de coëfficiënt van de toegevoegde variabele (t_h) en de toename van de R^2 die deze toevoeging oplevert[7)]:

7) Zie Theil (1971, blz. 125).

$$R^2 - R_h^2 = \frac{(1 - R^2)\, t_h^2}{n - k}$$

Het betekent dat, indien de t-waarde van de coëfficiënt van een variabele in absolute waarde gelijk aan 1 is, de gecorrigeerde $\bar{R}^2$ bij het toevoegen van die variabele niet verandert[8]. Is de t-waarde kleiner dan 1, dan wordt de $\bar{R}^2$ kleiner, en bij een t-waarde groter dan 1 neemt de $\bar{R}^2$ toe. Uit dit oogpunt is het dus aanbevelenswaardig om bij de beoordeling van regressieresultaten naast een t-waarde van 2 ook een t-waarde in de buurt van 1 of groter als interessante waarde te beschouwen.

Een dergelijke aanbeveling is eveneens te vinden bij Merkies (1972) die het voorspellingsinterval als maatstaf bij de kwaliteitsbeoordeling van een model gebruikt. Hij bewijst (zie pag. 76) dat de geschatte variantie van de residuen niet verandert indien de t-waarde van de coëfficiënt van de toegevoegde variabele gelijk aan 1 is. Deze stelling impliceert overigens onmiddellijk de hierboven beschreven eigenschap voor de gecorrigeerde $\bar{R}^2$. Immers, $\Sigma y^2 / (n-1)$ verandert niet bij toevoeging van een verklarende variabele en indien de schatting van de variantie $\Sigma e^2 / (n-k)$ ook niet verandert, blijft de $\bar{R}^2$ dus gelijk. Merkies (zie pag. 52) verkiest vanuit zijn gezichtspunt van het voorspellingsinterval de gecorrigeerde $\bar{R}^2$ als selectiecriterium boven de ongecorrigeerde R^2. De $\bar{R}^2$ (of R^2) vormt overigens slechts één van de vele mogelijke selectiecriteria, die bij de procedures om de beste verzameling verklarende variabelen voor een regressievergelijking te kiezen gebruikt worden[9]. In economisch onderzoek is het nut van dergelijke mechanische selectieprocedures nogal beperkt, aangezien de economische theorie vaak dwingend voorschrijft welke variabelen in de regressievergelijking dienen te worden opgenomen. Het gaat er dan veeleer om hoeveel een bepaalde variabele in de regressie tot de verklaring bijdraagt.

Bijdragen van de verklarende variabelen

Voor deze toerekening van de afzonderlijke bijdragen van de verklarende variabelen aan de R^2 (of $\bar{R}^2$) bestaat geen ondubbelzinnige oplossing. Theil (1971, blz. 181) propageert een methode waarbij het verschil $R^2 - R_h^2$ voor iedere verklarende variabele de bijdrage aangeeft die deze variabele levert in de totale verklaring van de variatie in Y. In het licht van de voorgaande exercities kan men zich bij deze methode van Theil afvragen of het niet beter is om, wanneer men de R^2 voor vrijheidsgraden corrigeert, deze vrijheidsgradencorrectie direct in de afzonderlijke bijdragen op te nemen. In dat geval benoemt men dus het verschil $\bar{R}^2 - \bar{R}_h^2$ als de bijdrage van de desbetreffende verklarende variabele tot de $\bar{R}^2$ en kan het voorkomen (bij een t-waarde kleiner dan 1) dat deze bijdrage negatief is.

Een nadeel van deze laatste werkwijze is echter dat de bijdrage van een bepaalde verklarende variabele ook groter dan 1 kan zijn. Dit gebeurt in het voorbeeld van Theil waar de vraag naar textiel verklaard wordt uit het inkomen en de prijs van textiel. Hier is de gecorrigeerde $\bar{R}_h^2$ (h: prijs voor textiel) van de regressie van de vraag naar textiel met het inkomen negatief, nl. -0,055, en de $\bar{R}^2$ van de volledige vergelijking groter dan 0,945 (= $1 + \bar{R}_h^2$), nl. 0,960, zodat de gecorrigeerde bijdrage van de prijs van textiel op 1,025 uitkomt. Een dergelijke anomalie kan er overigens op wijzen dat er iets met de specificatie aan de hand is.

Bij de methode van Theil tellen de afzonderlijke bijdragen alleen op tot de totale bijdrage, d.w.z. $\sum_h (R - R_h^2) = R^2$ indien alle verklarende variabelen loodrecht op elkaar staan en dus onderling niet gecorreleerd zijn. In de praktijk is dat zelden of nooit het geval en de afwijkingen kunnen dermate aanzienlijk zijn – in de hypotheekrentevergelijking van Den Butter, Dongelmans en Fase (1977) 0,42 bij een R^2 van 0,79 [10] – dat aan deze toerekeningsmethode slechts een bescheiden betekenis mag worden toegekend.

Besluit

In een aantal standaard regressieprogramma's[11] wordt naast de gewone determinatiecoëfficiënt de voor vrijheidsgraden gecorrigeerde determinatiecoëfficiënt afgedrukt en vaak rapporteren de onderzoekers alleen deze laatste grootheid bij hun uitkomsten. In dit artikel hebben we het gebruik om de R^2 voor vrijheidsgraden te corrigeren onder de loupe genomen. De R^2 kan gezien worden als de schatting in de waarnemingsperiode van een analoge parameter P^2 uit de populatie. Deze parameter wordt door de R^2 echter niet zuiver geschat. Hoewel zuiverheid een motief bij de vrijheidsgradencorrecties is levert ook de gecorrigeerde $\bar{R}^2$ geen zuivere schatting voor P^2 op. Barten heeft echter laten zien dat in het model waar de verklarende variabelen geen stochasten zijn, de verschillen klein zijn.

De R^2 wordt evenwel in de regressierekening slechts zelden met het statistisch oogmerk om hypothesen omtrent P^2 te toetsen gebruikt. De grootheid wordt meestal alleen berekend om een indruk te krijgen van de mate van aanpassing en ook wel om verschillende regressievergelijkingen met elkaar te vergelijken. In deze context corrigeert men voor vrijheidsgraden omdat een vergelijking met veel verklarende variabelen over het algemeen een betere aanpassingskwaliteit te zien geeft dan een vergelijking met een klein aantal verklarenden.

Het probleem is dat de R^2 eigenlijk geen goed criterium is om regressieresultaten met elkaar te vergelijken. Men mag alleen betekenis aan verschillen in R^2 toekennen indien de te verklaren variabele hetzelfde is en het dezelfde waarnemingen betreft. En dan nog dient men te bedenken dat wanneer een bepaalde vergelijking een hogere R^2 heeft dan een andere, de P^2 van die vergelijking nog niet hoger hoeft te zijn dan van de andere.

Het voordeel van de vrijheidsgradencorrectie van de R^2 ligt vooral bij die gevallen waarbij de te verklaren variabele hetzelfde is, maar waarbij de verklarende variabelen van de ene regressievergelijking een deelverzameling vormen van de verklarende variabelen van een andere regressievergelijking. We hebben laten zien dat de gecorrigeerde $\bar{R}^2$ toeneemt bij toevoeging van een extra verklarende variabele indien de t-waarde van de coëfficiënt van die variabele in absolute waarde groter dan 1 is. Anderzijds neemt

---

8) Bij de correctie van Barten geldt $\hat{R}^2 - \hat{R}_h^2 = 0$ indien $t_h \approx \sqrt{\frac{n-k}{n+k-\frac{5}{2}}}$

9) Zie bijv. A.A.M. Jansen (1978) en voor een overzicht M. Thompson (1978).

10) Bij de gecorrigeerde verschillen 0,39 met een $\bar{R}^2$ van 0,77.

11) Wij hebben deze correctie aangetroffen in de regressieprocedures van SPSS, in AUTO-ECON van de Wharton School of Finance and Commerce en in het transferfunctieprogramma van Charles R. Nelson Associates, Inc..

de $\bar{R}^2$ af indien de desbetreffende t-waarde kleiner dan 1 is. Deze stelling relativeert trouwens het belang dat men aan de $\bar{R}^2$ als vergelijkingsmaatstaf moet toekennen. Men kan immers net zo goed naar de t-waarden van de coëfficiënten kijken waarbij de waarde 1 als kritische grens in het oog gehouden dient te worden. Deze t-waarde van 1 speelt ook bij het door Merkies gepropageerde voorspellingsinterval als modelkeuzecriterium een rol.

Een nadeel van de vrijheidsgradencorrectie is dat de $\bar{R}^2$ negatief kan zijn terwijl de P^2 en ook de R^2 altijd tussen 0 en 1 liggen. Groot is dit nadeel echter niet daar men in het geval van een negatieve $\bar{R}^2$ toch te doen heeft met een regressie waarbij nauwelijks van enige verklaring sprake is.

Indien men bij de berekeningsmethode van Theil $\bar{R}^2 - \bar{R}^2_h$ i.p.v. $R^2 - R^2_h$ als bijdrage van de h^e verklarende variabele beschouwt kan deze bijdrage zowel negatief als ook groter dan 1 uitkomen. Dit laatste is een eigenaardige eigenschap van deze methode waarvan het nut overigens beperkt is.

De R^2 blijkt als maat voor de aanpassingskwaliteit van een regressievergelijking weinig statistische diepgang te bezitten. De deskundigen wisten dit overigens al lang en wat hen betreft mag vermelding van deze grootheid bij regressieresultaten achterwege blijven. Het is hier echter net zo als bij het pond of het ons, dat de kruideniers en groenteboeren volgens de Warenwet ook niet als gewichtsmaat op prijskaartjes mogen vermelden. Toch snapt iedere consument wat ermee bedoeld wordt. Zo valt het naar onze mening aan te bevelen, louter uit het oogpunt van traditie en gewoonte, de R^2 bij regressieresultaten te presenteren. En wanneer men dat nu eenmaal doet is het beter de R^2 voor vrijheidsgraden te corrigeren opdat de uitkomsten bij een beperkt aantal waarnemingen en/of een groot aantal verklarende variabelen niet geflatteerd worden. Ook al bezit de vrijheidsgradencorrectie eveneens weinig statistische diepgang.

Literatuur

T.W. Anderson, *An Introduction to multivariate statistical analysis,* Wiley, New York (1958).

A.P. Barten, *Note on unbiased estimation of the squared multiple correlation coefficient,* Statistica Neerlandica 16, 151–163 (1962).

F.A.G. den Butter, A.M. Dongelmans en M.M.G. Fase, *De vraag naar hypothecair krediet en de rentevorming op de hypotheekmarkt,* De Economist 125, 43–74 (1977).

W. Driehuis, *Fluctuations and growth in a near full employment economy,* Rotterdam University Press (1972).

L. de Haan en E. Taconis-Haantjes, *Asymptotic properties of a correlation coefficient type statistic connected with the general linear model,* Journal of Econometrics 7, nr. 1, 15–21 (1978).

A.A.M. Jansen, *Zoeken naar een geschikt regressiemodel,* V.V.S.-bulletin, jaargang 11, juli/augustus 1978, 16–22 (1978).

J. Koerts en A.P.J. Abrahamse, *On the theory and application of the general linear model,* Rotterdam University Press (1969).

J.H.C. Lisman, *Vrijheidsgraden in de Statistiek,* E.S.B. 31–8–1977, 841–843 (1977).

A.H.Q.M. Merkies, *Selection of models by forecasting intervals,* Reidel, Dordrecht (1972).

E. Pedhazur, *Analytic Methods in Studies of Educational Effects,* Review of Research in Education, 3, 243–287 (1975).

H. Theil, *Principles of Econometrics,* North-Holland, Amsterdam (1971).

M. Thompson, *Selection of variables in multiple regression,* part I, II, International Statistcal Review, 46, 3–19, 129–146 (1978).